平易な表現への言い換えに必要なテキスト修正処理*

藤田篤† 乾健太郎† 松本裕治†
† 奈良先端科学技術大学院大学情報科学研究科
{atsush-f,inui,matsu}@is.aist-nara.ac.jp

1 はじめに

われわれは，テキストの読解を支援する技術として，平易な表現への言い換え（テキスト簡単化）の実現に取り組んでいる [5]．(1) のように受身文を能動文に言い換えたり，(2) のように普段あまり用いない単語をなじみのある単語に言い換えることで，テキストの読みやすさを向上させようというものである[1]．

(1) s. 米国の多くの**地域が悪天候に見舞われた**。
t. **悪天候が**米国の多くの**地域を見舞った**。

(2) s. 激しい自動車戦争に進む**公算**が大きい。
t. 激しい自動車戦争に進む**可能性**が大きい。

われわれが開発を進めている読解支援システムの概要を図 1 に示す．このシステムは，ユーザが任意のテキスト（Web 上の文書など）中の理解できない部分を指示すると，(i) その部分テキストに潜む言語的な難しさをユーザの可読性のモデルに照らして特定し，(ii) その難しさを解消するように言い換える，というものである．このような，テキストが可読性などの評価基準を満たしているかどうかの評価と言い換えの生成を切り離した設計は，評価基準を取り換えることでユーザ・目的による違いを吸収できる，汎用的な枠組となる．

図 1 のシステムの核となる言い換えを実現するには，「言い換え知識」を獲得するための実験環境が必要となる．そこで，われわれは，さまざまな言い換えの事例研究を支援するための言い換えエンジン KURA を開発している [9]．KURA は，言い換えと同じく意味を変えない言語変換である機械翻訳の技術を範に，また，単一言語内の変換における特徴を考慮して，構文的トランスファ（構造変換）方式を採用している．この方式は，テキストの一部をそれと語彙・構文的に等価な表現に置き換えるものである．われわれは，語彙・構文的な言い換えの事例研究を通じて言い換えに必要な知識を収集し，KURA 上に実装している[2]．

ところで，トランスファ方式で単純にテキストの一部を置き換えると，さまざまな不適格性が生じてしまう．たとえば，(3) では，置き換えた単語の意味が文脈によって元の単語の意味とは変わっているし，(4) では，言い換えた箇所と周辺の表現との統語的つながりが不自然になっている．

(3) r. `N1(同概念語:N2) => N2`
s. 文語体、しかも難解な言葉が**随所**にある。
t.*... 難解な言葉が**各地**にある。

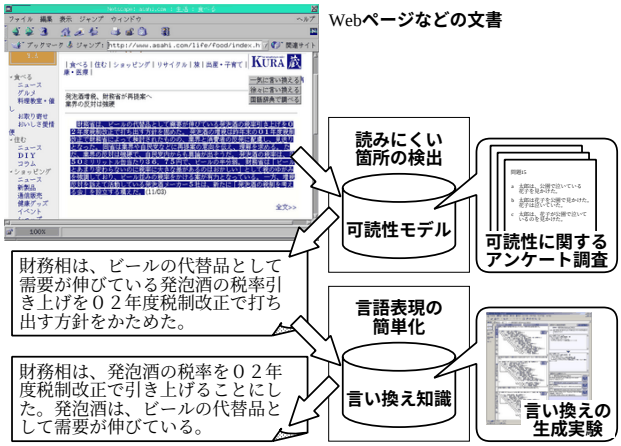

図 1: テキスト簡単化に基づく読解支援システム

(4) r. `N しか V ない => N だけ V`
s. あとは、ルイ・ヴィトンカップの**優勝艇しか出場できない**アメリカズカップでのレースが待っているだけ。
t.*あとは、ルイ・ヴィトンカップの**優勝艇だけ出場でき**アメリカズカップでのレースが待っているだけ。

言い換え後のテキストに含まれるこのような不適格性は，読解支援を考える上で致命的な場合が少なくない．したがって，テキスト中の不適格性を検出・解消する修正処理の実現が技術的課題となる．本稿では，このような修正処理の実現方法を提案し，現在実装を進めているモデルについて述べる．

2 言い換えに必要なテキスト修正処理

あらかじめ，トランスファと本稿で焦点を当てる修正処理を次のように定義しておく．

トランスファ: テキスト中のある適格な表現をそれと同義の別の表現に変換し，言い換えを生成する処理

修正処理: （トランスファに関係なく）ある不適格な表現を修正，または棄却する処理

トランスファによって生じる問題には，統語レベルから意味・談話レベルまで，性質の異なるさまざまなものがある．ここで，さまざまなトランスファ知識を用いて生成した言い換えに対して汎用的に作用する修正処理機構があれば，機械的に獲得，あるいは人手で記述した新規のトランスファ規則を導入する際に，人手で洗練するコストを軽減することができる．

言い換えに必要だと考えられる修正処理のうちいくつかは，すでに個別の言い換えの事例研究のなかで示

---

* Analysis and Modularization of Text Revision Processes for Text Simplification
FUJITA Atsushi†, INUI Kentaro† and MATSUMOTO Yuji†
† Graduate School of Information Science, Nara Institute of Science and Technology
{atsush-f,inui,matsu}@is.aist-nara.ac.jp

[1] 本稿の例中，r. は言い換えパタン，s.，t. は各々言い換え前後のテキスト，規則中の N，V はそれぞれ名詞句，動詞句を指す．

[2] http://cl.aist-nara.ac.jp/lab/kura/doc/

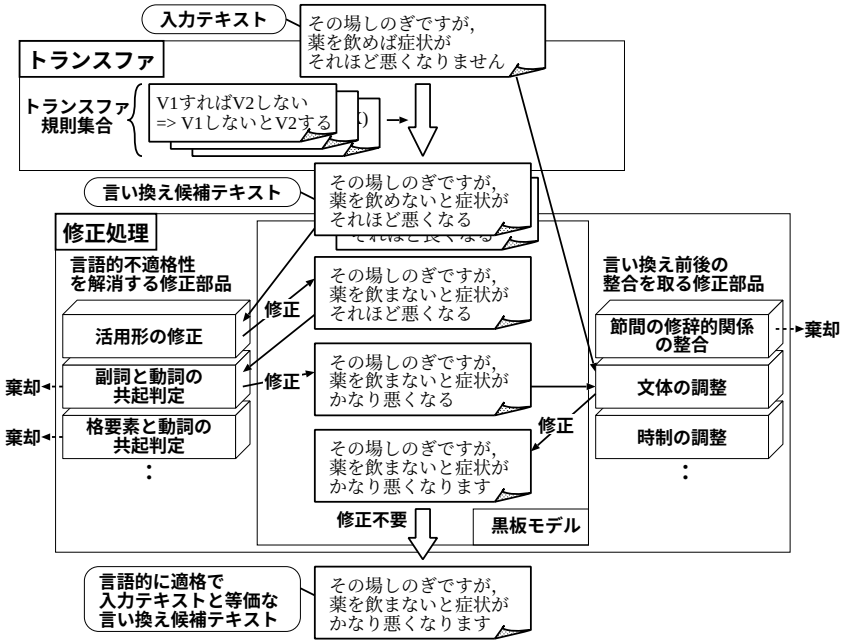

図 2: トランスファと修正による言い換えの生成モデル

表 1: トランスファ後に必要な修正処理の種類と分布

修正処理の種類	必要となる頻度
《語の活用形修正》	303
《機能語の連接修正》	78
《格助詞の補完》	8
《重複する格の整合》	11
《節内の格要素と動詞の整合》	**148**
《修飾語と動詞の整合》	2
《内容語の指示的意味の等価性評価》	168
《モダリティの持つ意味の等価性評価》	22
《時間情報の整合》	6
《文体の整合》	1
《語順の整理》	34
《主題・陳述構造の整合》	22
《節間，文間の修辞的関係の整合》	8
その他	115

されている．ただし，あらゆる語彙・構文的言い換えに必要な修正処理は網羅されておらず，再利用も容易ではない．そこで，われわれは，さまざまな種類のトランスファ知識を用いた言い換えの生成，分析を通じて，言い換えに必要な修正処理を可能な限り分解し，その種類と分布を調査した [1]．新聞記事中の文に手持ちの約 8,000 個のトランスファ規則を適用して生成した 630 事例に対し，修正すべき項目を列挙した結果，表 1 に示す修正処理の必要性と，その各々の分布が得られた．各修正処理についての説明は紙面の都合上割愛するが，詳細は文献 [1] を参照されたい．

3 節内の格要素と動詞の共起判定

言い換えに必要なテキスト修正処理のうち，(4.t) に対する活用形の修正は，たいていの場合に必要な統語的処理である．また，(3.t) に対する内容語の指示的意味の等価性評価も,「意味を変えないような変換」という位置付けの言い換えにおいて必須の修正処理である．同様に，(5) のように，ある節内の単語を別の単語に置き換えた場合，あるいは (6) のように態交替，動詞交替によって格要素と格助詞が変化した場合に，言い換えた節内の動詞と格要素の名詞の共起が不適格になってしまうことがたびたびある．共起が不適格な中には，(6.t) → (6.t') のように修正できる事例もあるが，予備調査の限りでは，修正できず棄却するしかない事例がほとんどであった．したがって，本稿では，不適格な共起を検出し棄却するというタスクに取り組む．

(5) s. 資質と能力を持った「個人」が世代や**国境を超えた**ネットワークで結ばれる。
t.*資質と能力を持った「個人」が世代や**国境を上回った**ネットワークで結ばれる。

(6) r. `N1` に `N2` が `V` される => `N1` が `N2` を `V` する
s. ゼネコン**問題に終止符が打たれる**。
t.*ゼネコン**問題が終止符を打つ**。
t'. ゼネコン問題に終止符**を**打つ。

3.1 動詞と名詞の共起を扱う関連研究

語彙レベルの（各単語に依存した）知識は，たとえば格フレーム辞書のような形で整備することが考えられる．動詞の下位範疇と選択制限の記述については NTT の日本語語彙大系 [4] がよく知られているが，言い換えにおける単語間の細かな用法の違いを捉えるには粗い．コーパスを用いて格フレーム辞書を構築・拡充する試みもある [2, 6] が，単語と単語の組み合わせまで考慮すると，やはり規模が小さい．

Utsuro ら [11]，Miyata ら [8] は，ある名詞がある動詞の下位範疇を満たすか否かを最大エントロピー法，ベイジアンネットワークといった確率モデルで推定し，動詞の多義性解消に応用している．これらに代表されるように，言語解析タスクの多くは，可能な解釈の中から一つの解を選択するというものである．

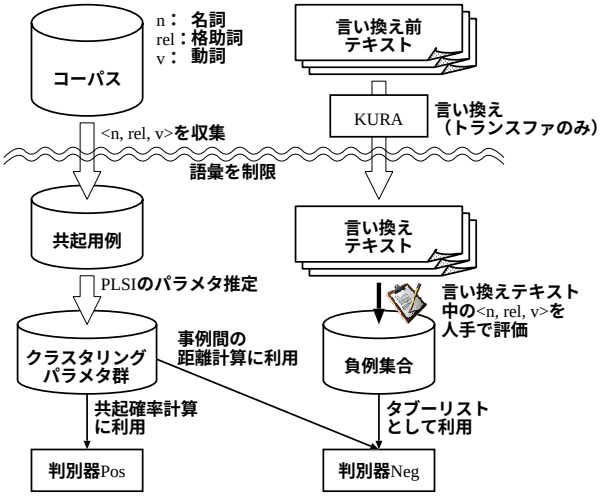

図 3: 2 つの判別器とそれに用いる知識の構築手順

これに対し，われわれの扱うタスクは，名詞と動詞の共起それ自身の正しさを評価することが求められる生成の問題である．棄却，あるいは修正といった処理も考えうるため，問題はより複雑である．

3.2 統計ベース教師なし学習と用例ベース教師あり学習を混合した判別モデル

評価対象は，トランスファによって置換，あるいは移動された単語を含む節である．ここで，節内のあらゆる 3 つ組 $\langle$ 格要素（今回は名詞のみ）n，格助詞 rel，動詞 $v\rangle$ の共起の適格さに基づいて，節全体の共起の適格性を判定することを考える．すなわち，$\langle n, rel, v\rangle$ を，適格かそうでないかの 2 値に分類する問題に帰着させる．

個々の単語に依存した問題であるため，統計的アプローチをとる．利用できる知識としては，コーパスから獲得できる大規模な共起用例（正例）がある．2 値分類問題であるため負例も用いる必要があるが，これは自然界には存在しないので人手で作成する．Utsuro らはシソーラスを利用したが，われわれは，人手で作成された知識を取り入れることを避ける．なぜならば，(2)，(3)，(5) のように，トランスファの時点ですでに，シソーラス中の同概念語を用いており，その同概念語間の用法の違い（正確には周辺の語との共起の適格さ）こそが捉えたいものだからである．

大規模な正例に基づいて $\langle n, rel, v\rangle$ の「正例らしさ」を見積もる．尤度を見積もるモデルにはさまざまなものがあるが，ここでは，単語の共起を潜在的な意味からの同時発生とみなし，その度合いを確率的に推定する，PLSI(Probabilistic Semantic Latent Indexing)[3] を用いる．このモデルを用いると，$\langle n, rel, v\rangle$ の共起確率は，次式で与えられる．

$$P(\langle n, rel, v\rangle) \stackrel{\text{def}}{=} \sum_{z\in Z} P(n|z)P(\langle rel, v\rangle|z)P(z)$$

式中のパラメタ $P(z), P(n|z), P(\langle rel, v\rangle|z)$ は EM アルゴリズムによって推定できる [3]．上式で与えられる共起確率は，あらゆる n，$\langle rel, v\rangle$ を考えたときに，どれだけ共起しやすいかというものであるが，本タスクでは n，および $\langle rel, v\rangle$ は既知であるので，その結び付きのよさ，すなわち「正例らしさ」は，確率そのものではなく，相互情報量，Dice 係数などによって見積もる．

少数の用例に基づいて「負例らしさ」を見積もる手法としては，最近隣検索法（Nearest Neighbor; NN）を用いる．NN では，距離を計算するために用例に何らかの素性を与える必要がある．この素性は，先に述べたように人手では付与せず，用例が PLSI の隠れクラス $z \in Z$ に帰属する確率の分布 $P(z|\langle n, rel, v\rangle)$ を用いる．これは，PLSI で得られるパラメタにベイズの定理を施すことで求められる．任意の用例 $\langle n, rel, v\rangle$ 間の距離は，Jensen-Shannon divergence（以下 JS）[7] で与える．JS は，Kullback-Leibler divergence に確率 0 の素性（正確には確率変数）への耐性と，対称性を持たせたものである．「負例らしさ」は，人手で作成した負例，すなわちタブーリストに基づき，上位 k 個の最近隣用例の JS の平均によって定量化する．

以上をまとめると，われわれの正負例の判別モデルは図 3 のようになる．このモデルは，PLSI で与えられる共起確率に基づく判別器 Pos，および人手で作成した負例に基づく判別器 Neg の 2 つによって $\langle n, rel, v\rangle$ の適格性を判定する．判別に必要な統計量および事例は，以下の手順で構築した．

1. 新聞記事 19 年分[3]を CaboCha [4] で係り受け解析し，3 つ組 $\langle n, rel, v\rangle$ を抽出した[5]．
2. 機械学習の高速化のために語彙を制限する．今回は，のべ 2000 回以上出現した名詞 3,365 語，動詞 2,516 語を採用した．また，格助詞は “が”，“を”，“に”，“で”，“へ”，“から”，“より” の 7 つとした．1. で得た 3 つ組のうち，3,628,345 組がこの制限語彙で表現されており，$\langle rel, v\rangle$ の異なりは 16,899 種，n と $\langle rel, v\rangle$ の共起行列要素充填率は 6.38%であった．
3. PLSI 学習パッケージ[6] を用いて，確率的パラメタを推定した．隠れ変数の個数 $|Z|$ は 100 とした．
4. 負の共起用例は言い換えによって自動生成したテキストから人手で収集する．ランダムに収集した 4,095 件の言い換え事例中，言い換えによって生成，移動された語を含み，かつその節内のすべての名詞，動詞が 2. の制限語彙に含まれているのは 809 事例であった．この中の $\langle n, rel, v\rangle$ の適格性を人手で判別した結果，正例 624 と，負例 185 を得た．

3.3 評価と考察

人手で正負のラベルを付与した 809 事例を用い，負例を検出する実験を行った．判別器 Pos による判別結果，判別器 Neg に基づく判別結果，および 2 つの判別器の判別結果を混合した結果を各々図 4，図 5，図 6 に示す．後者 2 つについては，人手で評価した負例を学習に用いるため，5 分割交差検定によって精度を算出している．なお，判別結果の混合方法としては，さまざまなパラメトリックな手法があるが，今回は負例らしさの論理和を用いた．すなわち，いずれかの判別器が負例とみなした場合に全体として負例と判定する．

[3] 毎日新聞 9 年分，日経新聞 10 年分，のべ 25,061,504 文.
[4] http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/
[5] のべ 53,157,450 組，異なり 7,993,331 組.
[6] http://cl.aist-nara.ac.jp/~taku-ku/software/plsi/

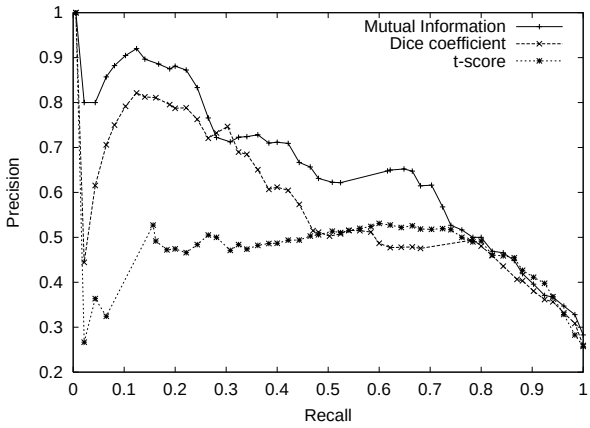

図 4: 判別器 Pos による負例検出の RP 曲線

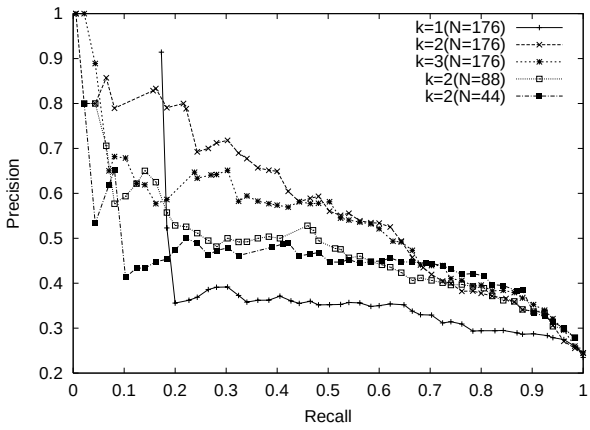

図 5: 判別器 Neg による負例検出の RP 曲線

Pos だけによる判別結果に比べ，Neg の判別結果を加えることで，わずかではあるが精度の向上が見られた．また，負例の学習量 N を増やすにしたがって精度が向上することが確認できたため，残る課題は，いかに効率的に負例を収集するかということになる．k-NN における k の変化に対しては，用例がそれほど多くはないため，$k = 3, 4, 5$ でほぼ集束した．

今回は兄弟格要素との共起の適格性までは判定していないが，実際は，(7) のように，各 $\langle n, rel, v \rangle$ の共起が適格でも兄弟の共起が不適格になる場合もあるため，モデルに組み込んでいくべきである．これについては，鳥澤が，兄弟格要素の組み合わせも捉えることができるモデルを示し，2 つの名詞の共起からそれらを結ぶ動詞を推定するタスク [10] などについて報告しているので参考になる．

(7) s. 二十歳代の**夫婦が**当時三歳の**長男に**十分な食事を**与えず**、...
t.* 当時三歳の**長男が**二十歳代の**夫婦から**十分な食事を**取らず**、...

4 おわりに

本稿では，読解支援の実現方法の一つとして平易な表現への言い換えに着目し，言い換えを生成する際に必須となる，テキスト修正処理の実現方法について述べた．そして，モジュール化を進めている修正処理の

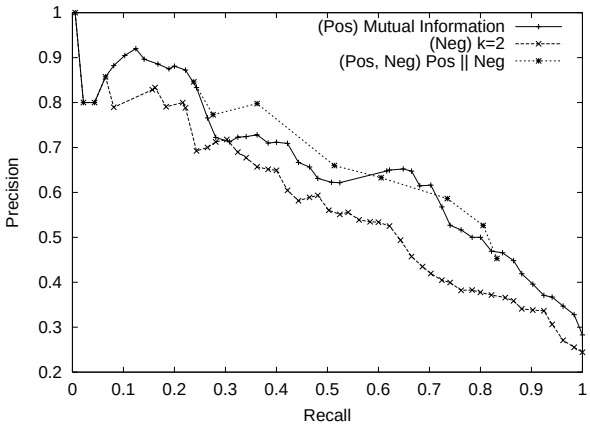

図 6: 判別器の混合による負例検出の RP 曲線

うち，名詞と動詞の共起の不適格性を検出するための統計モデルについて詳述した．現状では，われわれのモデルが頑健な言い換えの生成，および読解支援に十分貢献できているとは言明できないが，問題を切り分けることで，モジュール化の容易さ，部分問題の解決能力向上を示すことができた．今後も，各修正モジュールの精緻化，および未実装の修正モジュールの実現方法について考察する予定である．

謝辞 機械学習手法を導入するにあたり，さまざまな点で御助言下さいました奈良先端大の高村大也氏，持橋大地氏，工藤拓氏に感謝致します．

参 考 文 献

[1] 藤田篤, 乾健太郎. 語彙的言い換えに必要な知識の部品化. 情報処理学会自然言語処理研究会予稿集, NL-149-5, pp. 31–38, 2002.

[2] Sanae Fujita and Francis Bond. A method of adding new entries to a valency dictionary by exploiting existing lexical resources. In *Proc. of the 9th TMI*, pp. 42–52, 2002.

[3] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd SIGIR*, pp. 50–57, 1999.

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦（編）. 日本語語彙大系: CD-ROM 版. 岩波書店, 1997.

[5] 乾健太郎. 読解支援を目的とするテキスト簡単化の実現に向けて — 課題と方法論 —. 電子情報通信学会思考と言語研究会予稿集, TL2001-8, pp. 51–58, 2001.

[6] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol. 9, No. 1, pp. 3–19, 2002.

[7] Lillian Lee. Measures of distributional similarity. In *Proc. of the 37th ACL*, pp. 25–32, 1999.

[8] Takashi Miyata, Takehito Utsuro, and Yuji Matsumoto. Bayesian network models of subcategorization and their MDL-based learning from corpus. In *Proc. of the 4th NLPRS*, pp. 321–326, 1997.

[9] Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, Atsushi Fujita, and Kentaro Inui. KURA: a transfer-based lexico-structural paraphrasing engine. In *Proc. of the 6th NLPRS Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 37–46, 2001.

[10] 鳥澤健太郎. 教師無し学習による名詞句の言い換え. 言語処理学会第 8 回年次大会発表論文集, pp. 323–326, 2002.

[11] Takehito Utsuro, Takashi Miyata, and Yuji Matsumoto. Maximum entropy model learning of subcategorization preference. In *Proc. of the 5th Workshop on VLC*, pp. 246–260, 1997.